A Work Project, presented as part of the requirements for the Award of a Master's degree in Economics from the Nova School of Business and Economics.

DETERMINANTS OF HOUSEHOLDS' CONSUMPTION IN PORTUGAL
A MACHINE LEARNING APPROACH

CATARINA VIEIRA NORO

Work project carried out under the supervision of:

Paulo M. M. Rodrigues

04-01-2021

Determinants of Households' Consumption in Portugal - a Machine Learning Approach

Catarina Noro

Abstract

Machine Learning has been widely adopted by researchers in several academic fields. Although at a slow pace, the field of economics has also started to acknowledge the possibilities of these algorithm based methods for complementing or even replace traditional Econometric approaches. This research aims to apply Machine Learning data-driven variable selection models for accessing the determinants of Portuguese households' consumption using the Household Finance and Consumption Survey. I found that LASSO Regression and Elastic Net have the best performance in this setting and that wealth related variables have the highest impact on households' consumption levels, followed by income, household's characteristics and debt and consumption credit.

JEL Classification: C45, C60, D12, E21, E27

Keywords: Machine Learning Algorithms, Feature Selection, LASSO Regression, Elastic Net, Boosting, Bagging, Consumption

1 Introduction

The fast paced developments in Machine Learning (ML) and artificial intelligence (AI) have created more efficient research methods, which are useful across several research fields. These improvements have allowed the processing of high-dimension and high-volume datasets, which require more computational power. The potential for using these methods in the field of Economics is still being studied, but there is evidence that ML can complement or even, replace regular econometric methods.

Hereupon, this work project aims to firstly provide a brief revision of the ML routines and methods and secondly apply them to an economic problem setting: identifying individual consumption determinants for Portuguese households. ML data-driven feature selection models are used, instead of the traditional intuitive economic feature selection method, to access their potential for research in Economics.

According to Pordata, in Portugal, household's consumption represented 68,2% of GDP and 80% of total consumption, during 2019. As such, from the expenditure view point, household consumption is, with a large margin, one of the main drivers of GDP in Portugal. Accordingly, the consumption level is one of the key economic indicators to use for inference about the Portuguese economy.

Through an analysis of households' consumption levels and habits, one can infer about aggregate consumption and, thus, the country's economic performance. As such, it is imperative to study how household's consumption behaves, not only for informative and academic research purposes, but also for policy decisions to be more effective, targeting the right people at the right time.

In 2018, the 3 items Portuguese households consumed the most were "Food, beverages and tobacco", "Housing and utilities" and "Transports and communication". They represented approximately, 27%, 24% and 21% of household's total consumption, respectively. At the same time, household's mean disposable income was, on average, 33 119 € yearly. However, consumption over exceeded the mean disposable income in 2.3%, meanings families were consuming more than earning.

Household's consumption behavior can be affected by several factors. In this work project,

I divide the set of variables into 5 different categories: ”Income”, ”Wealth”, ”Household’s Characteristics”, ”Debt and Consumer Credit” and ”Expenses”. By studying and ranking their regression’s coefficients and importance, I analyze how they impact consumption. Since I am performing feature selection, this work project’s results could then be used for prediction models or even to further study causal effects in more traditional Econometric approaches.

In the second section of this work project, I present a brief explanation of what ML is, why and when it should be used and its potential for Economics. I present a Literature Review on the ML routines and the more useful models for economists. In the third section I explain how I chose the dataset and how was the curation process. In the fourth section, using the Python library Scikit-Learn (Pedregosa et al., 2011), I apply ML models in a feature selection problem setting. Household level data was retrieved from the Eurosystem Household Finance and Consumption Survey (HFCS) and, by using different variables’ sets I implemented 6 different ML models for accessing which ML model better worked in this problem setting. In the fifth section, I discuss and compare the variable selection results with the existing consumption Economic literature. Finally, the sixth section provides the conclusions of this work.

2 Literature Review

2.1 What is Machine Learning?

ML has been widely adopted by researchers in several academic fields. Although at a slow pace, the field of economics has also started to acknowledge the possibilities of these algorithm based methods for improvements in empirical work. As such, there is an increase of the ML methodological literature (Athey and Imbens, 2019).

ML is a part of artificial intelligence. It is the science that finds patterns and computes predictions through the development of models which are able to ”learn” from data. Its aim is to, taking into account a sample of data, generalize and perform predictions/actions in unseen data. It is mainly used for predictive purposes in classification or regression tasks, therefore, it has a great potential for Economics. There are recent developments into causal analysis using ML, however, the literature is not extensive. Some examples can be seen in Athey and Imbens

(2017) and Belloni et al. (2014). ML can complement or even, in specific scenarios, replace traditional econometric models.

2.2 Why should we use ML and in which context should we use it?

There are two main scenarios where the use of ML methods, rather than Econometric methods, are beneficial (Athey and Imbens, 2019).

Firstly, ML can contribute effectively to deal with Big Data. A dataset set is considered Big Data when there is a large number of sectional units or a large number of observations per sectional unit (Athey and Imbens, 2019). This type of dataset is being more common nowadays, given there are new digital data sources, which compile information at a much faster pace. The potential of using Big Data in the field of Economics is still being studied, but there is evidence that Machine Learning methods are more efficient at dealing with this unconventional amount of data than traditional Econometric methods (Mullainathan and Spiess, 2017). For Big Data applications in the Economics field see, e.g. Buono et al. (2017). In this work project, I will be using a dataset with more than 100 variables and perform variable selection. Although this dataset is still not considered Big Data, it could be beneficial to use ML models in this problem setting, since economic intuition can be inefficient when faced with this amount of variables.

Secondly, while ML is more concerned with prediction, Econometric models are built to summarize the causal effects between the variables and the outcome. In terms of recognizing complex structures and patterns in data, producing accurate predictions in out-of-sample contexts, ML is successful and prevents over-fitting problems. However, to date there has not been much evolution in ML methods regarding marginal effects and causal relations. To find these relations there is the need to consider estimators' assumptions and properties, such as consistency and efficiency. Therefore, ML methods usually do not provide theoretical results, only predictions (Mullainathan and Spiess, 2017). Nonetheless, for a review of the progress in ML regarding causal relations see, for instance, Athey and Imbens (2017).

2.3 Machine Learning approach vs Econometrics approach

In econometrics models are constructed recurring to economic theory. The model has a dependent variable and a set of covariates that, supposedly, explain the variation in the former. A random sample of the population of interest is used to measure how the variation of the covariates affects the variation in the dependent one, in other words, to find and quantify a causal relation. This process is done using objective functions (such as the sum of squared errors or a likelihood function) and provides the estimated marginal effects that ”best fit” the sample (Athey and Imbens, 2019).

On the contrary, the main objective of ML is to generate predictions and not to search for causality. As such, marginal effects are usually not considered. The models are composed of a target variable, also known as label, and a set of variables that describe the data, also known as features. To develop a prediction, the algorithm will ”learn” the patterns from a representative sample. After the learning process and tuning the model, the model can be used to present predictions from unseen data (Athey and Imbens, 2019).

2.4 Sample Splitting and Validation

Usually, the ML approach begins with splitting the sample into 3 different sets: training set, validation set and test set. The training set is created in order for the model to “learn” from the data. It will find the patterns and structure of that sample’s subset. However, it is likely that the model over-fits the training set, meaning that it will probably perform well at predicting in-sample, but will perform poorly in out-of-sample predictions (Varian, 2014).

To correct for over-fitness, there is first cross-validation and then regularization. Regularization will be covered in section 2.6.1. Cross-validation is performed through the tuning of hyperparameters, the parameters that can be changed in order for the model to perform better (e.g. the learning rate). It allows us to choose a well-performing combination of hyperparameters.

In cross-validation, after training the model, validation is then performed in the estimated model from the training process. Forecasts are made, followed by the calculation of forecast errors. Next, taking into account the errors, we perform an iterative search for the hyperparam-

eters that minimize the errors. The idea is to simulate an out-of-sample prediction, but using the results to continuously improve the model, through hyperparameter tuning (Varian, 2014).

Finally, after cross-validation, we use the test set to evaluate the model's out-of-sample performance (Gu et al., 2020).This evaluation can be done using different metrics, depending on whether it is a classification or a regression problem.

2.5 Supervised vs Unsupervised Learning

There are 3 types of different ML approaches: unsupervised learning, supervised learning and reinforcement learning. This literature review will not focus on reinforcement learning, since it is not a suitable approach for this analysis.

In supervised learning, the training set is accompanied by labels for each observation. Therefore, the predictions are computed based on the label and on the observable features. In contrast, in unsupervised learning the data set does not have labels. As such, unsupervised learning aims to find "hidden" patterns in the data. There are different tasks we can perform with unsupervised learning, for example we can split the data into clusters (groups with similar characteristics) or perform dimensionality reduction (simplifying the data structure without losing much information).

In this work project, I will be using supervised learning since I have the label for my dataset. I want to find the determinants of households' consumption, as such, the used label will be "Consumption on goods and services".

2.6 Relevant Supervised Models

2.6.1 Regularized Linear Regression

The simple linear regression is estimated using Ordinary Least Squares (OLS). In OLS, the parameters are chosen to minimize the sum of the squared residuals. Usually, this model is the baseline typically used for comparison with more complex models to highlight their contributions. The linear regression model is indicated to search for causal relations between 2 variables but, in a ML setting, it is commonly used to make baseline predictions. This model does not

allow nonlinear effects and does not cover interactions between covariates.

However, when the number of regressors increases approaching the number of observations, the linear regression starts to over-fit the noise, producing inefficient and inconsistent estimates, due to the high number of estimations the model has to perform (Gu et al., 2020).

To correct over-fitness in a linear regression model there are two most common approaches: getting more data to train the model or reducing the number of features through Regularization. Getting more data to train the model is usually difficult and time consuming. Thus, reducing the number of features or the "importance" of some of them is considered a better approach. However, in a dataset with a large number of features this task can be hard to perform. As such, the literature often uses penalizing methods for regularization and shrinkage, such as LASSO, Ridge and Elastic Net.

The Ridge regression is a linear regression with a L2-norm penalty. This regression is mostly used when data suffers from multicollinearity problems. Instead of using OLS, Ridge minimizes a penalized version of the sum of squared residuals. The following formula represents how the coefficients for the Ridge regression are computed (see Friedman et al. (2001)):

$$\hat{\beta}^{Ridge} = \underset{\beta}{\operatorname{argmin}}\{\sum_{t=1}^{N}(y_t - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \gamma \sum_{j=1}^{p} \beta_j^2\}$$

The penalty is proportional to the sum of squares of the coefficients. By adding a certain degree of bias to the regression coefficients, the standard errors are reduced since the bias tackles the overfitting. The chosen hyperparameter γ is the penalty and will make the coefficients shrink towards zero but never reaching it. The higher the γ, the higher the shrinkage degree (Friedman et al., 2001). Therefore, this model does not exclude any feature from the set of available features, and as such it is considered as a shrinkage method but not a variable selection method (Gu et al., 2020).

On the other hand, although the LASSO regression is also a variation of the linear regression model, it imposes an L1-norm penalty instead of an L2. It adds a penalty term that is proportional to the sum of the absolute values of the coefficients. The following formula represents how the coefficients for the LASSO regression are calculated (see Friedman et al. (2001)):

$$\hat{\beta}^{\text{LASSO}} = \underset{\beta}{\text{argmin}}\{\frac{1}{2}\sum_{\text{t}=1}^{\text{N}}(\text{y}_\text{i} - \beta_0 - \sum_{\text{j}=1}^{\text{p}}\text{x}_\text{ij}\beta_\text{j})^2 + \gamma\sum_{\text{j}=1}^{\text{p}}|\beta_\text{j}|\}$$

Contrarily to the Ridge regression, some features are not considered in this model, given that their weights can shrink to 0. As such, it is considered as a variable selection method.

Finally, Elastic Net is a combination of both Ridge and LASSO. The penalty of absolute values and the penalty of the squared values are used in the same model. The following formula represents the Elastic Net penalty (see Friedman et al. (2001)):

$$\gamma\sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$$

2.6.2 Regression Trees

A Regression Tree is constructed by an iterative process that continuously splits the data into smaller branches/leafs. In a first step, all the observations on the training set are in the same branch. Then, in the first iteration this data is split into two other branches, each branch representing a threshold for a specific feature. The split is made in a way that minimizes the sum of squared deviations from the mean in both branches (Breiman et al., 1984). This method is used in each branch iterating until it reaches the maximum depth chosen.

The tree development is done using solely the training/validation data. To perform out-of-sample predictions, the value of the predictor is compared to the values of the branches attributed during the training process. However, this process of sample splitting can easily lead to over-fitting problems if we over-split. Cross-validation can be used in order to find the best performing level of tree depth (Mullainathan and Spiess, 2017). The main advantages of using decision trees is that they capture interactions between features contrarily to the linear regression, they are easily interpretable since at each tree split the results are a sample average and, has such, the tree has a natural intuitive visualization (Gu et al., 2020).

However, as said before, decision trees are prone to over-fitting, and if they are too complex, they may not generalize well to perform out-of-sample predictions. Furthermore, if the data is unbalanced the tree can turn out to be biased, as such, the decision tree model is best fit for balanced datasets. Finally, decision trees are also very unstable, and one small change in the

configurations of the tree can give a total different output. This problem can be solved by using a combination of different decision trees, an ensemble method called Random Forest. Ensemble allows averaging the features across the different decision trees included. For an explanation about ensemble methods see section 2.8.

2.6.3 Artificial Neural Networks

To work with highly nonlinear systems or datasets with a very large number of features, Artificial Neural Networks are the most efficient model (Friedman et al., 2001). Currently they are used for complex ML settings, since they allow for a higher degree of flexibility due to the possibility of a high number of hidden layers, non-linearities and interactions between variables.

A neural network model is composed of nodes. There are three main types of nodes: the input layer, the hidden layers and the output layer. The nodes are connected to each other in a feedforward setting, meaning they do not form a circle and the information only flows forward beginning in the input layer, passing through the hidden layers and finishing in the output layer. The connections between the nodes have different weights and these weights are set during the training process. The stronger the connection the higher the weight. Each node receives input information from a previous node, performs a weighted sum and passes it to the next node using a non-linear function, also called activation function, until it reaches the output node (Efron and Hastie, 2016).

The model is trained with the back-propagation algorithm. Initially the weights are initialized randomly, the information is feedforward until the output node. When the information reaches the output node, the output error is calculated comparing the first prediction with the instance label. Then, the error is back-propagated from the output to the previous nodes. With this back-propagation method and the error information, the weights are adjusted. When reaching the input nodes, the process is repeated until the loss function converges (Efron and Hastie, 2016). Depending on the choice of the activation function, the model can tackle both classification and regression problems. Furthermore, the number of layers and nodes can be chosen and cross-validation is used in this task. However, training very complex models with a high number of layers and nodes is time-consuming and, due to the back-propagation algorithm, the

gradients are likely to ”explode” or ”vanish”, harming the training process (Gu et al., 2020). The “exploding” gradients can be corrected by applying a Batch Normalization algorithm.

2.7 Relevant Unsupervised Models

2.7.1 K-Means Clustering

Since in unsupervised learning there are no “labels” for the dataset, the task of this algorithm will be to find hidden patterns that divide the dataset into subsamples/clusters with similar characteristics.

Firstly, using the literature and economic interpretation, we set up an expected number of clusters, since there is no cross-validation method for this approach (Athey and Imbens, 2019). Then, the algorithm chooses centroids among the observations and divides the sample accordingly to the proximity of the observations to the centroids. An observation will be allocated to the cluster which has the closer centroid. After this first allocation, the centroids must be updated, averaging the features of each cluster (Athey and Imbens, 2019). With the average values of the clusters, through economic intuition one can perceive which ”profile”/”class” the clusters represent.

2.8 Ensemble: Bagging and Boosting

Another key feature that can be performed using ML is ensembling. Its purpose is to improve the performance of out-of-sample predictive models using combinations of different algorithms. There are two types of ensembling: Bagging and Boosting (Athey and Imbens, 2019).

The bagging algorithms compute a series of different models independently in parallel and then perform a model averaging. It is known that in general one model performs worse than a combination of different weighted models. This type of model averaging can be done with very different algorithms. The weights of each algorithm are computed by, using a test sample, minimizing the sum of the squared residuals. Since each model has its own pros and cons, a model averaging can improve out-of-sample predictions because it will take advantage of the strengths of each algorithm (Athey and Imbens, 2019). There are two different approaches of

bagging: one is to use very different models and average them, the other is to train the same framework on different random training sets with replacement of observations, also known as bootstrap sampling. One example of Bagging is the Random Forest model, a combination of different decision trees.

Boosting is when several models are built in a sequence. The next model is built taking into account the error of the previous model. After many iterations the improvements create a stronger final model (Athey and Imbens (2019) and Gu et al. (2020)).

3 Empirical Analysis

3.1 Data Cleaning

The dataset used in this work project was the Eurosystem Household Finance and Consumption Survey (HFCS). It is a European Union micro-level database, constructed by country level institutions such as national central banks and national statistical institutes in order to have harmonized data regarding euro zone household's finances and consumption.

The survey was conducted in three different parts. The first regards the household as a whole and is answered by a single household representative. This part has five different subparts: "real assets and their financing", "liabilities and credit constraints", "private businesses and financial assets", "intergenerational transfers and gifts" and "consumption/savings". The remaining two parts are conducted at an individual level, to every household member older than 16 years old. These individual surveys regard "employment", "future pension entitlements" and "labor-related income". The survey was then translated into a database with five different sets of variables: household's core and non-core variables, individual's core and non-core variables and derived variables. For the purposes of this work project, only household's core and derived variables were used, since they captured most of the variation in the target variable.

At the moment of this research, three years were available: 2010, 2013 and 2017. Since the target variable "HI0220: Amount spent on consumer goods and services" is not present in 2010, this year is not included in the analysis.

There are many null observations for "Don't know" or "No answer", as such, HFCS has

imputed values to cover the missing observations. For each missing observation there are five imputed values, to guarantee imputation uncertainty is accounted. This is translated into 5 different files for each set of variables. However, to avoid overfiting and data leakage, only one of these files was used in the execution of this project.

For Portugal, the household core dataset has 12131 observations and 1358 variables. These variables can be divided into 3 technical categories: survey's answers (675 variables), household's identification (8 variables) and flags (675). This dataset provides information about expenses, wealth and household's characteristics. However, information about income was missing, and according to economic theory, income plays an important role in determining an individual's consumption behavior. As such, 60 derived variables related to income and several other important indicators for the household were added to the household core dataset, adding up to a total of 1418 variables.

Through a brief analysis of the data, even after the imputation HFCS provides, there were a large number of null values (NaN). In ML models the dataset cannot present empty observations, therefore, it had to be curated.

In order to clean the missing values, all variables which presented a number of null observations equal or higher than 12 000 were dropped. This threshold was qualitatively chosen, since it is reasonable to delete features who are missing 99% of the observations. After this exclusion, there were 284 core variables. This means that 58% of the variables were missing 99% of the data.

After applying the quantitative threshold, a qualitative analysis was done to exclude variables that could cause noise. I decided to exclude 195 core variables, through qualitative reasoning. Thus, the final number of core and derived variables was 140, including the target variable. The list of all the used variables can be found in the Appendix.

However, even after deleting these variables, there were many null values that needed to be handled. These NaN values were caused by missing higher order answers. For example, if the first question was "Do you have sight accounts?" and the answer was "No", then the answer to the question "How much is the value of your sight accounts?" would be an NaN. To deal with this problem, I created a function using Python that would filter each null observation, taking

into account its flag and the nature of the question, and impute the number 0 in these cases or the number 2, when it was a categorical negative answer.

Having the dataset cleaned, the data should be standardized for comparison purposes. I used standarization since some of the variables have different units of measurement, for example "Age of reference person" is measured in years while "Employee Income" is measured in euros. The formula for standardization was:

$$z = \frac{x-u}{s}$$

where u is the mean and s is the standard deviation.

Having the data standardized, I examined the distribution of the target variable. As is possible to see from Table 1, there are 4 outliers. To reduce the noise and improve the model's fit, I chose to remove these outliers.

Bins	**Nr of Observations**
(-1.4, 3.8]	12007
(3.8, 9.1]	120
(9.1, 14.4]	3
(14.4, 19.6]	0
(19.6, 24.9]	1

Table 1: Data distribution for "Consumption on goods and services" - target variable

The final data frame had 12127 observations and 140 variables, excluding flags and identification variables.

3.2 Model Development

For all ML models I have used the Python library Scikit-Learn. All commands for the ML routine and models are available in this library.

The data was split into two parts: the training and the test set. The test set was set to 20% of the data and the training set to 80%. Throughout all models, the target variable is "HI0220" corresponding to the "Amount Spent on Consumer Goods and Services". The full list of the 141 features is provided in the appendix.

For this analysis, I divided the variables into 5 different categories: Income related, Wealth related, Expenses related, Household's characteristics and Debt and Consumer Credit related.

This category division is in line with Bouyon et al. (2015). He used disposable income of households, stock of consumer credit, nominal house prices and demographic trends to study what influences household consumption in the EU-28 at the aggregate level.

Firstly, I performed the models using all of these 5 categories of variables. However, since expenses are part of consumption they can be overshadowing the income and wealth effects. As such, I present here two model computations: with and without expenses.

I implemented 6 different models, from the 8 methods I mentioned in the Literature Review: LASSO Regression, Elastic Net, Decision Tree, Random Forest, AdaBoost and Gradient Tree Boosting. Ridge Regression was not used since it is a shrinking method and not a feature selection method. Artificial Neural Network is also not suited for this type of analysis since it is indicated for highly nonlinear complex systems. Although this dataset can have non-linearities, I expect the relation between the features and the consumption variable to be almost linear. Finally, the K-Means algorithm was also not used since it is suited for a problem setting where there are no labels. In this research, the label is the household's consumption.

3.2.1 LASSO Regression and Elastic Net

The first models to be applied are the LASSO Regression and the Elastic Net. As explained in section 2.6.1, LASSO Regression and Elastic Net are regularization methods which prevent overfitting by using penalties on the coefficients. Since the coefficient's weights in these models can shrink to 0, some of the coefficients vanish. This characteristic makes these models useful for feature selection and, therefore, capable of finding households' consumption determinants.

These two models have a hyperparameter that needs to be tuned: the alpha. The alpha corresponds to the penalty and the higher its value, the higher the shrinkage degree. To choose the best performing alpha, one should perform cross-validation. I performed cross-validation using 5-folds. The table below presents the performance for these 4 models and their best performing hyperparameter.

From Table 2, it is possible to see that LASSO regression and Elastic Net have very similar performances. However, the models which include the Expenses variables perform better than the models which do not include them, the difference is around 23 percentage points in R^2.

Model	Alpha	R^2
LASSO with Expenses	0.008	73.07
LASSO without Expenses	0.008	49.70
Elastic Net with Expenses	0.015	73.06
Elastic Net without Expenses	0.017	49.74

Table 2: LASSO Regression and Elastic Net - Hyperparameters and model performance

3.2.2 Decision Tree

The Decision Tree model is not linear and captures interactions between variables, unlike the LASSO Regression and the Elastic Net. Due to its more complex structure, the Decision Tree has more than one hyperparameter to tune. As such, a simple cross-validation does not provide the necessary information. For tuning the decision tree a Grid Search Cross-Validation was used. This method performs all the possible combinations of a set of pre-defined parameters. For this analysis I tuned maximum depth, the maximum number of features and the minimum sample split.

The table below represents the best hyperparameters combination for the model with and without Expenses. As seen before, the model with Expenses has a higher R^2 than without Expenses. However, the best performing Decision Tree has a lower performance than both LASSO Regression and Elastic Net. The cause of this can be the linear relation between consumption and the features. Furthermore, the low R^2 could also mean there is instability in the model. In order to solve for instability, we can use ensemble methods, creating a Random Forest, or using boosting algorithms as AdaBoost and Gradient Tree Boosting as seen in the next subsections.

Model	Max Depth	Max Features	Min Sample Split	R^2
DT with Expenses	6	105	5	57.95
DT without Expenses	4	70	5	40.49

Table 3: Decision Tree - Hyperparameters and model performance

3.2.3 Random Forest

Since the Decision Tree was not performing well as expected, I chose to run a Random Forest. As explained before, a Random Forest is a combination of several Decision Trees. Through the

averaging of features across different estimators, the model will generate a more stable output, more robust to small changes in its configuration and input.

Since this model is very dense and complex a Grid-Search Cross-Validation is very time consuming. As such, I opted for a Randomized Search Cross-Validation, which, instead of going through all the possible combinations of the hyperparameters, it does a random search across the possible combinations of a pre-defined set of parameters. For this model I tuned maximum depth, maximum features, minimum sample split, number of estimators, and whether to use bootstrap or not.

The table below shows some of the tuned hyperparameters and the model performance.

Model	**Nr Estimators**	**Max Depth**	**Max Features**	R^2
RF with Expenses	180	70	133	71.45
RF without Expenses	230	20	$\sqrt{121}$	50.30

Table 4: Random Forest - Hyperparameters and model performance

As in the previous estimators, the model which includes Expenses has a higher R^2 than the model which does not include Expenses. Furthermore, the best performing Random Forest has a higher R^2 than the Decision Tree, as expected. This ensemble algorithm yields an improvement of around 13 p.p. in R^2 from the Decision Tree.

3.2.4 AdaBoost

AdaBoost is a boosting algorithm. Instead of combining several estimators, it processes the error's information of several weak performing models in order to create a more powerful single final estimator. As base estimator for AdaBoost, I used the best performing Decision Tree. For tuning the Learning Rate and the Number of Estimators, Grid-Search Cross-Validation was used.

The table below shows the results for the cross-validation.

Model	**Learning Rate**	**Nr Estimators**	R^2
AdaBoost DT with Expenses	0.1	100	68.71
AdaBoost DT without Expenses	0.1	100	41.03

Table 5: AdaBoost - Hyperparameters and model performance

Like the previous models, including Expenses yields a higher R^2. Furthermore, although using the boosting algorithm improves the performance of the initial Decision Tree by 10 p.p. , AdaBoost performance is worse than the Random Forest's bagging method.

3.2.5 Gradient Boosting Regression Tree

Since AdaBoost did not perform as well as expected, I chose to implement another boosting algorithm, best suited for Regression Trees: the Gradient Boosting Regression Tree (GBRT). Since this model is dense and complex as the Random Forest, a Randomized Search Cross-Validation was used for tuning the hyperparameters. For this model I tuned maximum depth, maximum features, minimum sample split, number of estimators, and the learning rate.

The table below presents the results after cross-validation.

Model	**Nr Estimators**	**Max Depth**	**Max Features**	R^2
GBRT with Expenses	180	60	$\sqrt{133}$	67.33
GBRT without Expenses	180	60	$\sqrt{121}$	43.32

Table 6: Gradient Boosting Regression Tree - Hyperparameters and model performance

From the results, the Gradient Boost yields a lower R^2 than the AdaBoost when using Expenses, but the R^2 is higher than AdaBoost when the expenses variables are not used. As such, the best performing ensemble method is still bagging making the Random Forest the best performing model when not includinh Expenses.

4 Results Discussion

4.1 Models with Expenses variables

From the previous section, it is clear that the linear models over perform the non-linear models when including expenses. A reason for this could be the fact that, since expenses are part of consumption, they are more correlated with the target variable. LASSO Regression has the highest R^2. The model explains approximately 73.07% of consumption in goods and services variation. However, Elastic Net also performs similarly, explaining 73.06% of consumption variation, when using expenses.

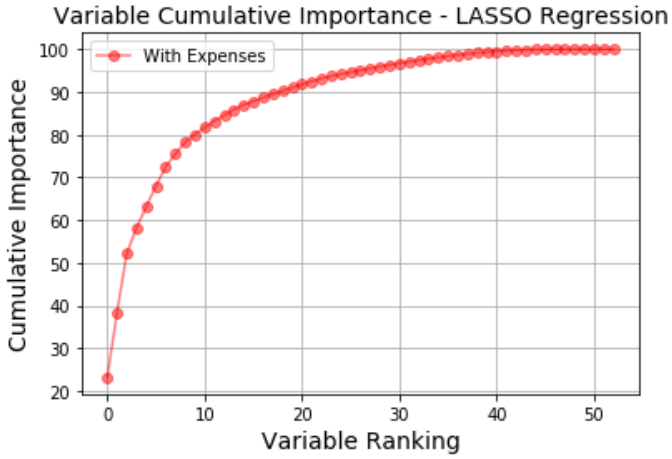

Figure 1: Cumulative Importance for LASSO regression - with Expenses

The LASSO regression selected a total of 53 different variables and shrunk the coefficient of 80 variables towards 0. By ranking the absolute value of the regression coefficients, it is possible to see which variables have the highest impact on consumption. The higher the absolute coefficient, the higher the impact. Figure 1 shows that the 10 variables with larger absolute coefficients explain approximately 80% of the total model variance when including Expenses.

Figure 2 presents the coefficients for the top 10 variables with highest impact on consumption. Note that these coefficients cannot be directly interpreted, because the data was standardized during the data curation process. Although the coefficients lost their economic interpretation, it is still possible to use the coefficients to perform variable ranking.

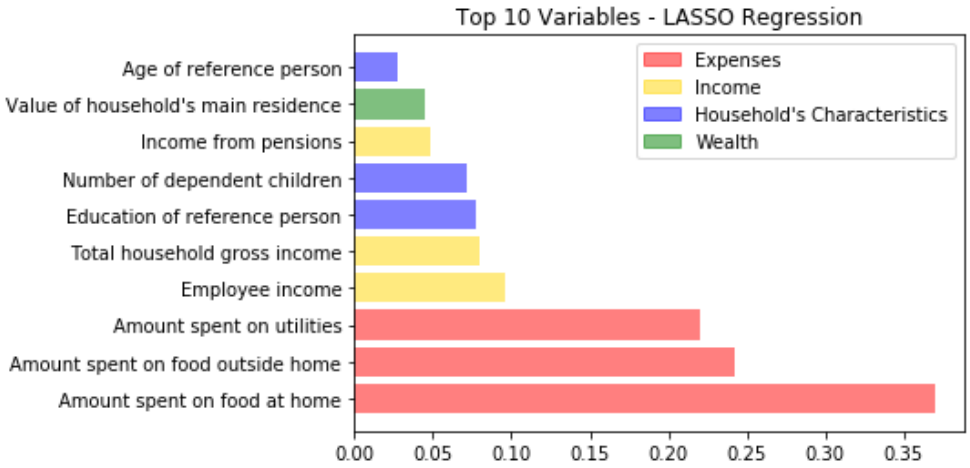

Figure 2: Highest Variables' Coefficients - LASSO Regression with Expenses

From these 10 variables, Expenses have the highest impact on consumption since they have the highest coefficients. This behavior was expected because “Amount spent on food at home”, ”Amount spent on food outside home” and ”Amount spent on utilities” are all components of household’s consumption. Variations in the amount spent on food at home have more impact in consumption than meals taken outside home and the consumption of utilities, indicating that food at home represents a higher portion of consumption. This is in line with what the Pordata aggregate data portraits: “Food, beverages and tobacco” represent 20% of household’s total expenses while “Restaurants and Hotels” represent only 13,9%. However, these results show that the “Amount spent on utilities” represents a smaller proportion of variation in consumption, than the “Amount spent on good outside home”.

After Expenses, the variables which capture more consumption variance are Income related. ”Employee Income” regards all the income the household obtained through salaries and other employment income. In turn, ”Total Household Income” is the sum of income from employment, self-employment, real estate, financial assets, pensions, regular social transfers, regular private transfers and other sources. ”Income from pensions” corresponds to the income that is received during retirement. Variations in employee income have more effect in consumption than variations in the total household income. One explanation for this could be the fact that total household income is more volatile, since it incorporates self-employment and financial assets income. Thus, since households are expecting this variation, its variation does not impact consumption as the more stable employee income. Since the model is linear, it could also indicate employee income has a more linear relation with consumption, than total household income.

Household’s characteristics also play an important role in household’s consumption. The higher the level of education of the reference person (individual answering household’s related questions), the higher the consumption level. This relationship could be derived from the fact that more educated people have, on average, more income and wealth to spend. For more information on the relationship between consumption and education levels see, for instance, Michael (1975).

According to the regression results, more children in the household also yield a higher

consumption level. More dependent children can mean that the household has also a larger number of members and, as such, needs to have a higher level of consumption. Apps and Rees (2001) estimate that the consumption of two dependent children represents between 23% and 34% of a traditional household's total consumption.

Finally, the older the reference person, the higher the consumption. This can happen because age is very correlated with years of experience. When the individual has more experience, its salary is expected to be higher. With higher employee income, consumption also tends to increase, as seen before. As such, age has a positive relation and a strong impact on consumption. Williams (2009) indicated an increase in wages of 60% after 30 years of experience in the UK, while Topel (1991) suggested an increase of 25% after 10 tenure years in the USA.

Surprisingly, there is only one Wealth variable and no Debt and Consumer Credit variable in the top 10 most important variables for consumption. The value of the main residence is the measure of wealth that most impacts consumption. Households which possess a more valuable residence have higher levels of consumption. Attanasio et al. (2005) and Attanasio and Weber (2010) describe how house prices impact consumption and prove that there is a positive relation between these two variables.

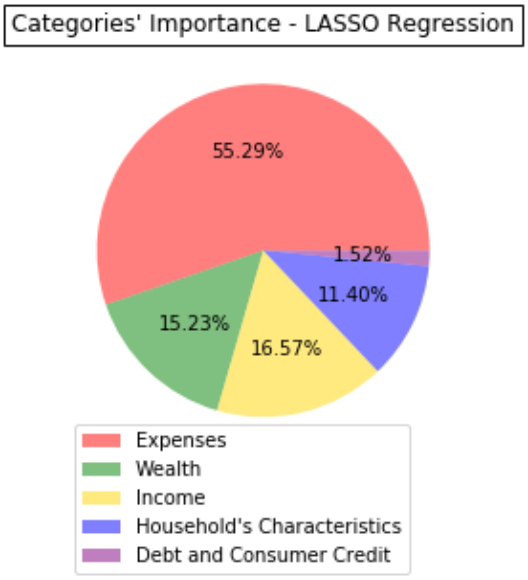

Figure 3: Categories' importance as share of total importance - LASSO Regression with Expenses

Another key information that the results show is presented in the Figure 3. When summing all the importance of the chosen variables, the impact from the Wealth variables is slightly

higher than the impact of the Income variables. This can mean that the Wealth effect on consumption is higher than the Income effect. Household's Characteristics and Debt and Consumer Credit are the categories which yield the lowest impact on consumption.

Since the expenses are overshadowing income and wealth, to further see the effect of these two categories, the next subsection presents the results without Expenses.

4.2 Models without Expenses Variables

The model with the best performance without Expenses is the Random Forest (RF). The RF explains approximately 50% of consumption's variance. This model selected 121 variables, meaning it incorporated all the input variables. This would not be ideal for our problem setting, however, this happens since, by nature, the main purpose of a RF is not feature selection, it is prediction. Nevertheless, since RF can also eliminate variables and provide feature importance, it is also used in the feature selection settings.

RF does not provide model coefficients but in turn, provides a measure of importance. The sum of all variables' importance adds up to 100. The higher the importance, the higher the variable's impact on the household's consumption. Figure 4 shows that the 10 variables with the highest importance explain around 60% of the model's variance.

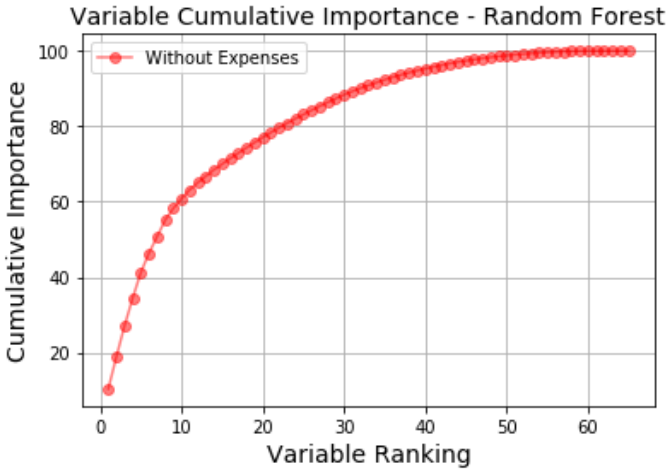

Figure 4: Cumulative importance for Random Forest - without Expenses

For illustrative purposes, I will analyze the 10 features with highest importance. Figure 5 shows "Total Household Gross Income" having the highest importance, explaining 13.5% of the

model's variation. In second place, in terms of feature importance, is the "Employee Income". Like the LASSO Regression with Expenses, "Total Household Income" and "Employee Income" have a higher importance than Wealth related variables and Household's Characteristics, in the top 10 variables with higher impact. At a first glance, this could mean that the Income category has a higher impact than Wealth. However, when summing all the variables' importance, Wealth accounts for 51.08% of the model's variation, while Income only accounts for 33.93%, as seen in Figure 6. This means that, although certain single Income variables have a high impact on consumption, Wealth as a whole explains more variance in consumption.

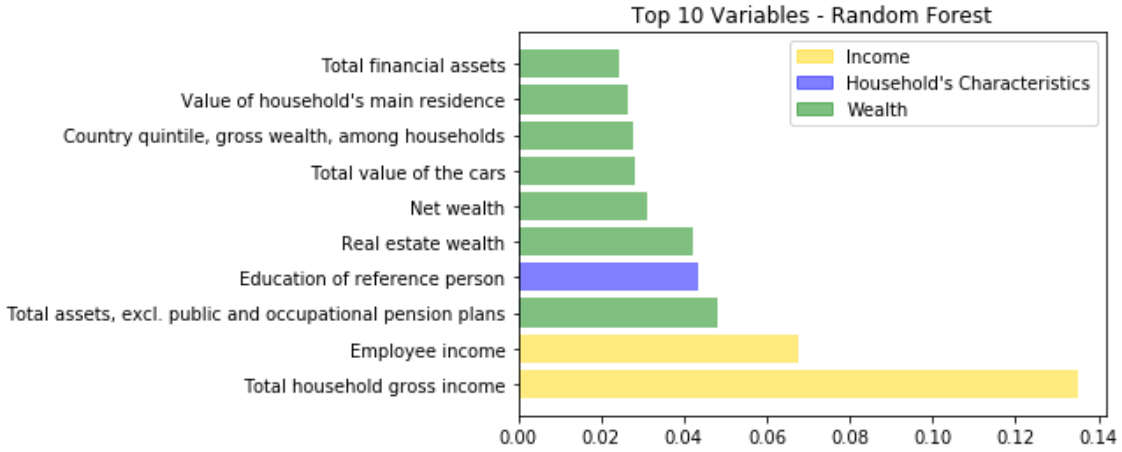

Figure 5: Highest Variables' importance - Random Forest without Expenses

In this model's output, the only household characteristic in the 10 most important variables is the "Education of the reference person". Again, families where the reference person has a higher education level have a higher level of consumption.

Seven out of the top 10 variables with higher importance are Wealth related. Contrarily to the LASSO Regression without Expenses, the Wealth variable with more impact on consumption is "Total assets". Households with more assets also tend to consume more. This change in the ranking of the variables can be due to the fact that the RF also captures non-linear effects.

In Figure 6, without expenses, all other variables capture a higher share of importance. The category which increases its share the most is Wealth, representing 46,16% of the model's variation, followed by Income and Household's Characteristics, capturing 33.93% and 14.99%, respectively. Debt and Consumer Credit continues to be the least important category, only representing, approximately, 4,93% of the model's variation.

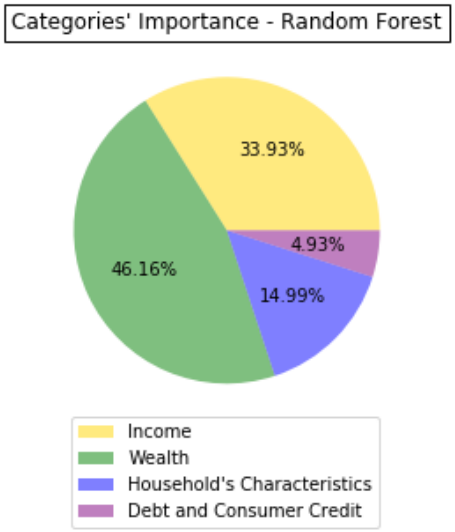

Figure 6: Categories' importance as share of total importance - Random Forest without Expenses

5 Conclusion

For this work project, I studied the application of ML models in a Economics feature selection problem setting. After providing a brief ML literature review, I used Portuguese household level data as input for the ML models, in order to understand which features have an higher impact on households' consumption. The analysis was made by single variables and by category.

By performing hyperparameter tuning using different methods for cross-validation, I found that the LASSO Regression and the Elastic Net had the best performance overall. These models had the best performances while including Expenses related variables, explaining approximately 70% of consumption variation. When excluding the Expenses category, the best performing model was the Random Forest, explaining 50.3% of consumption. Nevertheless, Elastic Net and Lasso Regression explained 49.74% and 49.70%, respectively.

While ranking by importance the variables of the best performing models with and without Expenses, I found that Expenses were overshadowing the other categories. Since expenses are part of consumption, they explained 55.29% of the model's variation. When not including the Expenses category, all the other categories gain in importance. However, Wealth is the one which gains the most, representing 46.16% of the Random Forest's variation, followed by Income (33.9%), Household's characteristics (14.99%) and, finally, by Debt and Consumer Credit (4.93%).

While developing this work project I understood there are advantages and disadvantages of using ML methods for feature selection. On the one hand, due its computational power, ML allows the examination of hundreds of variables at the same time. When faced with a dataset with a large amount of variables, these methods are useful for an economist. Nowadays, with the appearance of new alternative sources of data, like Big Data, ML can become an important asset for economists in a feature selection setting. For developing ML models, Scikit-Learn Python library is a suitable tool. It is of simple application, has all the models that were explained in this project, and provides the user with all the necessary documentation for a better understanding.

On the other hand, ML models have their drawbacks. Firstly, they do not allow the models to run with null observations. It is time consuming to curate all the NaNs, and if curation is not well performed, it can easily introduce bias in the dataset and, consequently, in the model's results. Secondly, if variables are presented in different scales, data needs to be standardized and the interpretability of coefficients is lost, meaning causal inference is not possible. Lastly, in a complex dataset, sometimes it is still necessary to perform a qualitative pre-selection. This means it is not possible to rely solely on a data-driven approaches if the aim is to have meaningful results. Economic intuition and reasoning are, therefore, mandatory.

All in all, ML models can be useful for economists in a feature selection setting. They have their drawbacks, however, they can be a good complement to the traditional Econometrics methods. With larger datasets appearing and, thus, an increased demand for more computational power, ML will be a necessary tool for future economic analysis.

References

Apps, P. and R. Rees (2001). Household production, full consumption and the costs of children. *Labour Economics 8*(6), 621–648.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*, 685–725.

Attanasio, O., L. Blow, R. Hamilton, and A. Leicester (2005). Consumption, house prices and expectations. *Available at SSRN 824744*.

Attanasio, O. P. and G. Weber (2010). Consumption and saving: models of intertemporal allocation and their implications for public policy. *Journal of Economic literature 48*(3), 693–751.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Bouyon, S. et al. (2015). Household final consumption in the eu: The key driver for a sustainable recovery? Technical report, Centre for European Policy Studies.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Buono, D., G. L. Mazzi, G. Kapetanios, M. Marcellino, and F. Papailias (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators 1*(2017), 93–145.

Efron, B. and T. Hastie (2016). *Computer age statistical inference*, Volume 5. Cambridge University Press.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Michael, R. T. (1975). Education and consumption. In *Education, income, and human behavior*, pp. 233–252. NBER.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Topel, R. (1991). Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of political Economy 99*(1), 145–176.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Williams, N. (2009). Seniority, experience, and wages in the uk. *Labour Economics 16*(3), 272–283.

6 Appendix

6.1 Variables Used

VARIABLES	DESCRIPTION
HB0400	is rent paid for partially owned household main residence
HB2300	monthly amount paid as rent
HC0100	household has a leasing contract
HC0110	monthly leasing payments
HI0100	amount spent on food at home
HI0200	amount spent on food outside home
HI0210	amount spent on utilities
HI0220	amount spent on consumer goods and services
HI0300	Makes private transfers to individuals out of household/charities (y/n)?
HI0310	amount given as private transfers per month
HI0500	comparison of last 12 months expenses with average
HI0600	last 12 month expenses were below/above income
HG0100	received income from public transfers
HG0110	gross income from regular social transfers
HG0200	received income from regular private transfers
HG0210	income from regular private transfers
HG0300	received income from real estate proprerty
HG0310	gross rental income from real estate property
HG0400	received income from financial investments
HG0410	gross income from financial investments
HG0500	received income from private business other than self-employment
HG0510	gross income from private business other than self-employment
HG0600	received income from other income sources
HG0610	gross income from other income sources
HG0700	is income 'normal' in reference period
HG0800	future income expectations
DI1100	DI1100 Employee income
DI1200	Self-employment income
DI1300	Rental income from real estate property
DI1410	Income from financial assets, gross of interest payments
DI1400	Income from financial investments
DI1420	Income from private business other than self-employment
DI1510	Income from public pensions
DI1520	Income from occupational and private pensions
DI1500	Income from pensions

VARIABLES	DESCRIPTION
DI1610	Unemployment benefits
DI1620	Other social transfers
DI2000	Total household gross income
DI1100i	Has employee income
DI1200i	Has self-employment income
DI1400i	Has income from financial investments
DI1410i	Has income from financial assets, gross of interest payments
DI1420i	Has income from private business other than self-employment
DI1500i	Has income from pensions
DI1510i	Has income from public pensions
DI1520i	Has income from occupational and private pensions
DI1600i	Has income from regular social transfers (except pensions)
DI1610i	Has income from unemployment benefits
DI1620i	Has income from other social transfers
DI1700i	Has income from regular private transfers
DI1800	Income from other sources
DI1800i	Has income from other sources
DI2000eq	Equivalised household gross income
DITOP10	Country top 10% total gross income
DITOP10eq	Country top 10% gross equivalised income
HB0100	size of household main residence
HB0100_B	size of household main residence brackets
HB0500	% of ownership of household main residence
HB1000	mortgages or loans using HMR as collateral
HB1010	number of mortgages or loans using HMR as collateral
HB2400	household owns other properties than HMR
HB2410	number of properties other than household main residence
HB4300	ownership of cars
HB4310	number of cars
HB4400	total value of the cars
HB4500	ownership of other vehicles
HB4600	total value of other vehicles
HB4700	ownership of other valuables
HB4710	value of other valuables
HB4800	purchase of vehicles
HB4810	price of purchased vehicles
HC0200	household has credit line or overdraft
HC0210	household has outstanding credit line/overdraft balance
HC0220	amount of outstanding credit line/overdraft balance
HC0300	household has a credit card
HC0310	household has outstanding balance on credit cards
HC0330	has private loans
HC0340	how many private loans
HC0400	has any non-collateralised loans
HC1400	not applying for credit due to perceived credit constraints

VARIABLES	DESCRIPTION
HD0100	investments in businesses not publicly traded
HD0200	investments in self-employment businesses
HD0210	how many self-employment businesses
HD1100	household owns sight accounts
HD1110	value of sight accounts
HD1200	household owns savings accounts
HD1210	value of saving accounts
HD1300	household owns investments in mutual funds
HD1400	household owns bonds
HD1420	market value of bonds
HD1500	household owns publicly traded shares
HD1510	value of publicly traded shares
HD1520	any shares issued by foreign companies
HD1600	household owns managed accounts
HD1700	does anyone owe money to household
HD1800	investment attitudes
HD1900	any other financial assets
HH0100	any substantial gift or inheritance received
HH0110	no of gifts/inheritances received
HH0201	gift/inheritance #1: year gift/inheritance received
HH0202	gift/inheritance #2: year gift/inheritance received
HH0203	gift/inheritance #3: year gift/inheritance received
HH0401	gift/inheritance #1: value
HH0402	gift/inheritance #2: value
HH0403	gift/inheritance #3: value
DA1110	Value of household's main residence
DA2109	Voluntary pension/whole life insurance
DA1400	Real estate wealth
DA2100	Total financial assets 1 (excl. public and occupational pension plans)
DA3001	Total assets, excl. public and occupational pension plans
DL1000	Total outstanding balance of household's liabilities
DN3001	Net wealth
DHAQ01	Country quintile, gross wealth, among households
DL1231i	Has private loans
DA2199	Other types of financial assets
DA2199i	Has other types of financial assets
DATOP10	Country top 10% gross wealth
DLTOP10	Country top 10% total liabilities
DNFPOS	Net financial position [Net financial wealth]
DNNLA	Net liquid assets
DNNLAi	Has net liquid assets
DNNLAratio	Net liquid assets as a fraction of annual gross income
DODNI	Net wealth to income ratio of indebted households
HB0200	how long have you been living in the household main residence
HB0300	main residence - tenure status
HB0600	way of acquiring property

VARIABLES	DESCRIPTION
HI0800	ability to get financial assistance from friends or relatives
DH0001	Number of household members
DH0006	Number of household members 16+
DH14P	Number of household members aged 14+
DHN013	Number of children in household (0-13)
DH0003	Number of economically active members in household
DH0004	Number of household members in employment
DHaged65plus	Household members aged 65 or more
DHchildrendependent	Number of dependent children
DHHST	Housing status
DHAGEH1	Age of reference person
DHEDUH1	Education of reference person
DHEMPH1	Main labour status of reference person